이학석사 학위논문

이진분류 문제에서 교차검증방법과 일반화근사교차검증방법의 비교

The Comparison of Cross-Validation and Generalized Approximate Cross-Validation For Binary Classification Problem

2018년 7월

서울대학교 대학원

통계학과

황 참 이

이진분류 문제에서 교차검증방법과 일반화근사교차검증방법의 비교

지도교수 김 용 대

이 논문을 이학석사 학위논문으로 제출함

2017년 12월

서울대학교 대학원

통계학과

황 참 이

황참이의 이학석사 학위논문을 인준함

2017년 12월

위 원 장	이 영 조	(인)
부위원장	김 용 대	(인)
위 원	임 요 한	(인)

국문초록

일반적으로 교차검증방법(Cross validation 이하 CV)은 다양한 통계적 방법론에 적용되어 모형의 적절성을 조율해주는 도구로 활용되었다. 또한 일반화교차검증방법(Generalized cross validation 이하 GCV)은 제곱 손실오차를 가지는 가우시안 벌점화 선형모형에서 계산을 효율적으로 줄이는 도구로 활용되었다. 즉, 선형모형에 벌점화가 있는 경우 근사의 과정을 거쳐, 교차검증방법을 통해 예측력이 높은 모형을 적절히 선택할 수 있다. 이러한 논리를 확장하여, 이진 반응변수를 갖는 벌점화 로지스틱모형에서 교차검증방법을 일반화한 일반화근사교차검증방법(Generalized Approximate Cross Validation 이하 GACV)을 살펴본다. 일반화근사교차검증오차는 비가우시안 자료에서 벌점화 로그 가능도 회귀모형의 평활모수(Smoothing parameter)를 추정하는 역할을 한다. 일반화근사교차검증오차는 테일러 전개를 통해 목적함수를 1차적으로 근사시키고 GCV에서 사용한 방법처럼 평활행렬 대각원소의 산술평균을 이용하여 한번 더 근사시켜 얻을 수 있다. 본문에서는 자료 분석을 이용하여 GACV로 얻은 평활모수의 성능이 기존 교차검증방법과 크게 다르지 않음을 보인다.

주요어 : 평활모수, 벌점화 로지스틱 회귀분석, 교차검증방법, 일반화근사교차검증방법

학 번 : 2016-20280

Contents

List of Tables

List of Figures

Chapter 1

서론

검증오차법(validation error)은 여러 모형의 예측력을 이용하여 최적의 모형을 선택하며 주어진 자료의 크기가 충분히 큰 경우 사용되는 방법이다. 주로 모형의 예측력에 초점을 맞춘 방법으로 알려져 있으며 이를 일반화한 것이 교차법검증법이다. 가령 자료를 서로 배반이 되도록 무작위의 k개 묶음(fold)으로 분할하여 각각의 검증오차를 구할 수 있으며 k개 검증오차의 산술평균을 흔히 k-묶음 교차확인오차(k-fold cross validation)라고 부른다. 이러한 방법을 극단적으로 사용하여 묶음의 갯수가 주어진 자료 수 n과 일치한다면 하나 남겨놓기 교차검증법(leaving-one-out cross validation)이라 부른다. 하지만 하나 남겨놓기 교차검증법은 교차검증법에 비하여 묶음의 갯수가 크기 때문에 편의-분산 절충에 의하여 검증오차의 기대 편의는 줄어들지만 분산이 큰 단점이 있다. 또한 선형모형과 같은 특수한 경우를 제외한다면, 개별 모형에 표본크기 $n-1$인 모형 적합을 n번 반복 적용해야하는 문제점이 한계로 지적된다.

지금까지 교차검증법의 경우 가령 로지스틱, LDA 등 다양한 통계 방법에 적용될 수 있기 때문에 꾸준히 교차검증법에 대한 근사가 이루어졌다. 대표

적으로 가우시안 선형모형에서 일반화 교차검증오차(GCV)와 같은 방법이 있다. 하지만, 단순 벌점화 선형모형이 아닌 비가우시안 비선형모형이라면 정형화된 검증방법이 마련되어 있지 않다. 음 로그가능도의 2차 근사를 통한 GCV(O'sullivan)[5]와 같은 방법, 불편 위험 추정치(Gu)[2] 등 다양한 방법이 사용되어 왔지만 예측오차에 대한 정확한 불편추정치를 얻는 획일화된 방법은 존재하지 않는다. 그러므로 예측오차의 불편추정치를 제시하는 여러가지 방법이 꾸준히 제시되어 왔으며 이러한 이유로 본문 제 2장에서는 $GACV(\lambda)$를 소개하고 이것으로 베르누이 자료에서 평활모수를 추정하였다. 그리고 제 3장은 실제 자료 분석을 바탕으로 비교를 통해 $CV(\lambda)$와 $GACV(\lambda)$ 성능이 크게 다르지 않음을 밝힌다. 마지막으로 제 4장에서 결론을 맺는다.

Chapter 2

분석방법론

이 장에서는 이진 반응변수에 대한 벌점화 로지스틱 회귀모형과 비교 분석에 사용된 교차검증방법을 간략하게 언급한다. 그리고 마지막으로 일반화 근사교차검증방법에 대해 자세히 다루도록 하였다.

2.1. 이진 반응변수에 대한 벌점화 로지스틱 회귀모형

선형 회귀분석 모형은 흔히 연속형 반응변수 y와 하나 또는 여러 개의 설명변수 사이의 선형적인 관계를 모형화한 통계적 분석방법을 의미한다. 만약 y가 질적변수, 특히 이진 반응변수라면, 로짓 연결함수를 활용하여 로지스틱 회귀분석을 주로 사용하며 또한 고차원 자료의 경우 안정적인 분석 결과를 얻기 위하여 벌점화 모형을 사용하므로, 이 두가지를 결합하여 이진 반응변수에 대한 벌점화 로지스틱 회귀모형[4]을 분석에 사용하였다.

2.2. 교차검증방법(CV)

앞으로의 논의를 위해 사용할 지수족에 대해 간략하게 다루도록 한다. 비가우시안 지수족을 갖는 자료구조에서, 벌점화 로그 가능도 평활 모수를 추정하는 방법에 대해 생각해보자. 여기서 y_i 는 독립인 관측치이고 다음의 지수족을 따르는 분포에서 추출된 임의표본이라 가정하자.

$$f(y_i, \psi(x_i), \phi) \equiv exp\{(y_i\psi(x_i) - \xi(\psi(x_i)))/d(\phi) + h(y_i, \phi)\} \tag{2.1}$$

여기에서 d, ξ, h는 주어져 있으며 $\xi(\cdot)$은 순볼록 함수이다. 흔히 $\psi(x_i)$를 표준모수라고 하며, 주된 목표는 $\psi(\cdot)$를 추정하는 것이다. 가령, Bernoulli라면

$$\begin{aligned} P(Y_i = y_i \mid x_i) &= p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \\ &= exp\{y_i\psi(x_i) - \log(1 + e^{\psi(x_i)})\} \end{aligned}$$

이므로 $\psi(x_i) = x_i^t\beta$, $\xi(x) = \log(1 + e^x)$ 이며 $h(y_i, \phi) = 0, d(\phi) = 1$ 이다. 모수적 GLM 모형에서는 $\psi(\cdot)$은 모수적 형태로 가정되지만 더 유연한 모형을 적합하기 위해, 스플라인 회귀 모형과 같은 방법이 적용될 수 있다. 가령 $\psi(\cdot)$를 재생커널힐버트공간의 원소로서 부드러운 함수 중 하나로 가정할 수 있다 [6]. 하지만 논의의 범위를 간소화하기 위해 $l(y_i, \psi(x_i)) = y_i\psi(x_i) - \xi(\psi(x_i))$로 한정하여 정의하기로한다. 그리고 $\psi(x_i) = x_i^t\beta$ 이고 $\xi(\psi(x_i)) = \log(1 + e^{\psi(x_i)})$ 라고 정의한다.

$\psi(\cdot)$의 벌점화 로그 가능도 추정치 $\psi_\lambda(\cdot)$는 식 (2.2)의 최소값으로 얻을 수 있다. 여기에서 평활모수 $\lambda \geq 0$을 만족한다.

$$\sum_{i=1}^{n} -l(y_i, \psi(x_i)) + \frac{n}{2}\lambda J(\beta) \tag{2.2}$$

즉, 음 로그 가능도 부분 $l \equiv -\sum_{i=1}^{n} l(y_i, \psi(x_i))$과 smoothness 조절하는 벌점항 $J(\beta)$ 사이를 조절하는 λ를 적절히 선택해야 한다.

선형 회귀모형에서 사용했던 교차검증방법 또는 일반화교차검증방법과 유사하게 로지스틱 벌점화 능형회귀 모형에서 사용할 교차검증오차는 다음과 같이 정의하였다[7].

$$CV(\lambda) \equiv \frac{1}{n}\sum_{i=1}^{n}[-y_i\psi_\lambda^{(-i)}(x_i) + \xi(\psi_\lambda(x_i))] \tag{2.3}$$

수식에서 사용한 $\psi_\lambda^{(-i)}(x_i)$는 i번째 관측 벡터를 제거하고 얻은 식 (2.2)의 최소값을 의미한다.

2.3. 일반화근사교차검증방법(GACV)

앞서 정의한 교차검증오차 $CV(\lambda)$를 근사하고 일반화시켜 벌점화 로지스틱 능형회귀 모형에 사용할 일반화근사교차검증오차를 직접 유도한다. 하지만 유도과정에서 사용한 $\psi_\lambda^{(-i)}(x_i)$은 하나의 공변량을 갖는 모형에서 반복 상태공간 알고리즘을 통해 $CV(\lambda)$를 구할 수 있으나[1], 일반적인 상황에서 사용되기에는 계산에 많은 시간이 소요된다. 따라서 일반화근사교차검증오차(이하 $GACV(\lambda)$ 오차)를 통해 위의 문제를 해결할 것이다.

$\mu_\lambda(x_i) = \frac{e^{x_i^t\beta_\lambda}}{1+e^{x_i^t\beta_\lambda}}$ 즉 $\mu_\lambda(x_i) = \xi'(\psi_\lambda(x_i))$라고 두고, 1차 테일러 급수를 이용하여 $CV(\lambda)$를 전개하는 $GACV(\lambda)$ 유도 과정은 다음과 같다.

식 (2.2)를 최소로하는 λ에 대하여 $l^\star \equiv -\sum_{i=1}^{n} l(y_i, \psi_\lambda(x_i)$라 하자. 그러면

$$\begin{aligned}
CV(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}[-y_i\psi_\lambda^{(-i)}(x_i) + \xi(\psi_\lambda(x_i))] \\
&= \frac{1}{n}\sum_{i=1}^{n}[-y_i\psi_\lambda(x_i) + \xi(\psi_\lambda(x_i))] + y_i[\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)] \\
&= \frac{1}{n}l^\star + \frac{1}{n}\sum_{i=1}^{n} y_i[\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)] \\
&= \frac{1}{n}l^\star + \frac{1}{n}\sum_{i=1}^{n} y_i \frac{\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)}{y_i - \mu_\lambda^{(-i)}(x_i)} \frac{y_i - \mu_\lambda(x_i)}{1 - \frac{\mu_\lambda(x_i) - \mu_\lambda^{(-i)}(x_i)}{y_i - \mu_\lambda^{(-i)}(x_i)}}
\end{aligned}$$

그리고 테일러 1차 근사를 통하여 아래의 식을 유추해 낼 수 있다.

$$\frac{\mu_\lambda(x_i) - \mu_\lambda^{(-i)}(x_i)}{y_i - \mu_\lambda^{(-i)}(x_i)} = \frac{\xi'(\psi_\lambda(x_i)) - \xi'(\psi_\lambda^{(-i)}(x_i))}{y_i - \mu_\lambda^{(-i)}(x_i)} \approx \xi''(\psi_\lambda(x_i)) \frac{\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)}{y_i - \mu_\lambda^{(-i)}(x_i)}$$

위에서 얻은 근사의 결과를 이용하여 $CV(\lambda)$를 근사시키면,

$$\begin{aligned}
CV(\lambda) &\approx \frac{1}{n}l^\star + \frac{1}{n}\sum_{i=1}^{n} y_i \frac{\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)}{y_i - \mu_\lambda^{(-i)}(x_i)} \frac{y_i - \mu_\lambda(x_i)}{1 - \xi''(\psi_\lambda(x_i)) \frac{\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)}{y_i - \mu_\lambda^{(-i)}(x_i)}} \\
&= \frac{1}{n}l^\star + \frac{1}{n}\sum_{i=1}^{n} y_i \frac{y_i - \mu_\lambda(x_i)}{\frac{y_i - \mu_\lambda^{(-i)}(x_i)}{\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)} - \xi''(\psi_\lambda(x_i))} \qquad (2.4)
\end{aligned}$$

를 얻게 되지만 식 (2.4)에서, $\frac{y_i - \mu_\lambda^{(-i)}(x_i)}{\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)}$를 계산해야하는 문제가 생기기 때문에, 마찬가지로 이것 역시 근사하는 방법이 필요하다.

먼저 *Leaving-out-one lemma*를 살펴보자. $-l(y_i, \psi(x_i)) = -y_i\psi(x_i)+\xi(\psi(x_i))$와 $Q_\lambda(\beta, \mathbf{y}) \equiv -\sum_{j=1}^{n} l(y_j, x_j^t\beta)+\frac{n\lambda}{2}J(\beta)$라고 하자. 여기에서 $J(\beta) \equiv||\ \beta\ ||_2^2$이고 $\omega_\lambda(i, z) \equiv \underset{\beta}{\operatorname{argmin}}\, Q_\lambda(\beta, \mathbf{z})$라 정의하자. 그러면

$$\omega_\lambda(i, \mu_\lambda^{(-i)}(x_i)) = \beta_\lambda^{(-i)}$$

임을 보일 수 있다. 참고로 $\mathbf{z} \equiv [y_1, \cdots, y_{i-1}, z, y_{i+1}, \cdots, y_n]^t = \mathbf{y} - y_ie_i + ze_i$라 두었고, $\beta_\lambda^{(-i)} \equiv \underset{\beta}{\operatorname{argmin}}\{-\sum_{j\neq i}^{n} l(y_j, x_j^t\beta)+\frac{n\lambda}{2}J(\beta)\}$ 라 정의하며, $\mu_\lambda^{(-i)}(x_i) = \frac{e^{x_i^t\beta_\lambda^{(-i)}}}{1+e^{x_i^t\beta_\lambda^{(-i)}}}$ 이다.

식 (2.4)를 근사하여 GACV를 유도하기 위해 간단하게 위의 정리를 증명하자. 우선 $\mathbf{y^{-i}} = [y_1, \cdots, y_{i-1}, \mu_\lambda^{(-i)}(x_i), y_{i+1}, \cdots, y_n]^t$라 정의하자. 그러면

$$-l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta) = -\mu_\lambda^{(-i)}(x_i)x_i^t\beta + \xi(x_i^t\beta)$$

이고 이것을 β에 대하여 미분하면,

$$\begin{aligned} -\frac{\partial}{\partial\beta} l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta) &= -\mu_\lambda^{(-i)}(x_i)x_i + \xi'(x_i^t\beta)x_i \\ -\frac{\partial^2}{\partial\beta\partial\beta^t} l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta) &= \xi''(x_i^t\beta)x_ix_i^t \end{aligned}$$

이다. 앞서 지수족에서 $\xi(\cdot)$은 순볼록함수이므로 $-\frac{\partial^2}{\partial\beta\partial\beta^t} l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta) \succcurlyeq 0$ 이고 로짓 연결함수는 일대일대응이므로 따라서,

$$\begin{aligned} -\frac{\partial}{\partial\beta} l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta) = 0 &\Leftrightarrow \xi'(x_i^t\beta) = \mu_\lambda^{(-i)}(x_i) \\ &\Leftrightarrow \beta = \beta_\lambda^{(-i)} \end{aligned}$$

을 결과로 얻을 수 있다.

결론적으로 $-l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta)$ 를 최소화 하는 β는 $\beta_\lambda^{(-i)}$이므로 임의의 $\beta \in \mathbb{R}^p$에 대하여,

$$-l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta) \geq -l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta_\lambda^{(-i)})$$

이므로

$$\begin{aligned} Q_\lambda(\beta, \mathbf{y}^{-i}) &= -l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta) - \sum_{j\neq i}^{n} l(y_j, x_j^t\beta) + \frac{n\lambda}{2}J(\beta) \\ &\geq -l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta_\lambda^{(-i)}) - \sum_{j\neq i}^{n} l(y_j, x_j^t\beta) + \frac{n\lambda}{2}J(\beta) \\ &\geq -l(\mu_\lambda^{(-i)}(x_i), x_i^t\beta_\lambda^{(-i)}) - \sum_{j\neq i}^{n} l(y_j, x_j^t\beta_\lambda^{(-i)}) + \frac{n\lambda}{2}J(\beta_\lambda^{(-i)}) \end{aligned}$$

이다. 따라서 *Leaving-out-one lemma*, $\omega_\lambda(i, \mu_\lambda^{(-i)}(x_i)) = \beta_\lambda^{(-i)}$, 를 증명하게 된다. 그러면 $(\beta_\lambda, \mathbf{y})$와 $(\beta_\lambda^{(-i)}, \mathbf{y^{(-i)}})$ 는 $Q_\lambda(\beta, \mathbf{z})$의 국소 최소 인자이므로,

$$\begin{aligned} \frac{\partial Q_\lambda(\beta, \mathbf{z})}{\partial\beta}(\beta_\lambda, \mathbf{y}) &= 0 \\ \frac{\partial Q_\lambda(\beta, \mathbf{z})}{\partial\beta}(\beta_\lambda^{(-i)}, \mathbf{y^{(-i)}}) &= 0 \end{aligned}$$

이다. 따라서, 1차 테일러 급수 전개를 이용하여 $(\beta_\lambda, \mathbf{Y})$에서 $\frac{\partial Q_\lambda(\beta, \mathbf{Z})}{\partial\beta}(\beta_\lambda^{(-i)}, \mathbf{Y^{(-i)}})$ 를 전개하면 다음의 식을 얻을 수 있다.

$$\begin{aligned} 0 &= \frac{\partial Q_\lambda(\beta, \mathbf{z})}{\partial\beta}(\beta_\lambda^{(-i)}, \mathbf{y^{(-i)}}) \\ &= \frac{\partial Q_\lambda(\beta, \mathbf{z})}{\partial\beta}(\beta_\lambda, \mathbf{y}) + \frac{\partial^2 Q_\lambda(\beta, \mathbf{z})}{\partial\beta\partial\beta^t}(\beta_\lambda^\star, \mathbf{y}^\star)(\beta_\lambda^{(-i)} - \beta_\lambda) \\ &\quad + \frac{\partial^2 Q_\lambda(\beta, \mathbf{z})}{\partial\beta\partial\mathbf{y}^t}(\beta_\lambda^\star, \mathbf{y}^\star)(\mathbf{y}^{(-i)} - \mathbf{y}) \end{aligned} \tag{2.5}$$

참고로 $(\beta_\lambda^\star, \mathbf{y}^\star)$은 $(\beta_\lambda, \mathbf{y})$와 $(\beta_\lambda^{(-i)}, \mathbf{y^{(-i)}})$ 사이에 존재한다.

그리고

$$\begin{aligned}\frac{\partial Q_\lambda(\beta,\mathbf{z})}{\partial\beta} &= \frac{\partial}{\partial\beta}\left\{-l(z,x_i^t\beta)-\sum_{j\neq i}^{n} l(y_j,x_j^t\beta)+\frac{n\lambda}{2}J(\beta)\right\}\\ &= \frac{\partial}{\partial\beta}\left\{-(\mathbf{y}-y_ie_i+ze_i)^tX\beta+\sum_{i=1}^{n}\log(1+e^{x_i^t\beta})+\frac{n\lambda}{2}\beta^t\beta\right\}\\ &= -X^t(\mathbf{y}-y_ie_i+ze_i)+\sum_{i=1}^{n}\left(\frac{e^{x_i^t\beta}}{1+e^{x_i^t\beta}}x_i\right)+n\lambda\beta\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 Q_\lambda(\beta,\mathbf{z})}{\partial\beta\partial\beta^t} &= \sum_{i=1}^{n}\left(\frac{e^{x_i^t\beta}}{(1+e^{x_i^t\beta})^2}x_ix_i^t\right)+n\lambda I_p\\ &\equiv W+n\lambda I_p\end{aligned}$$

를 각각 미분의 결과로 얻을 수 있다. 참고로 위의 수식에서

$$W = X^t\begin{bmatrix} p(x_1)(1-p(x_1)) & 0 & \cdots & 0\\ 0 & p(x_2)(1-p(x_2)) & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & \cdots & p(x_n)(1-p(x_n))\end{bmatrix}X$$

라고 정의하였으며, 축약하여 $W\equiv X^tPX$로 표현하였다. 그리고

$$\frac{\partial^2 Q_\lambda(\beta,\mathbf{z})}{\partial\beta\partial\mathbf{y}^t} = -X^t$$

역시 $\mathbf{y}$와 β로 목적함수를 미분한 결과 $-X^t$를 유도 할 수 있다. 여기에 각각 $X\beta_\lambda^{(-i)}=\psi_\lambda^{(-i)}$와 $X\beta_\lambda=\psi_\lambda$라고 두었으며, $W_\lambda^\star = X^tP_\lambda^\star X$이고, $W_\lambda = X^tP_\lambda X$

으로 표현하자. 따라서 1차 테일러 근사식 (2.5)은 식 (2.6)을 이용하여 정리할 수 있다.

$$
\begin{aligned}
&(W_\lambda^\star + n\lambda I_p)(\beta_\lambda^{(-i)} - \beta_\lambda) = X^t(\mathbf{Y}^{(-i)} - \mathbf{Y}) \\
\Leftrightarrow\ & X(\beta_\lambda^{(-i)} - \beta_\lambda) = X(W_\lambda^\star + n\lambda I_p)^{-1}X^t(\mathbf{Y}^{(-i)} - \mathbf{Y}) \\
\Leftrightarrow\ & \psi_\lambda - \psi_\lambda^{(-i)} = X(W_\lambda^\star + n\lambda I_p)^{-1}X^t(\mathbf{Y} - \mathbf{Y}^{(-i)})
\end{aligned}
$$

$$
\Leftrightarrow \begin{bmatrix} \psi_\lambda(x_1) - \psi_\lambda^{(-i)}(x_1) \\ \vdots \\ \psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i) \\ \vdots \\ \psi_\lambda(x_n) - \psi_\lambda^{(-i)}(x_n) \end{bmatrix} \approx X(W_\lambda^\star + n\lambda I_p)^{-1}X^t \begin{bmatrix} 0 \\ \vdots \\ y_i - \mu_\lambda^{(-i)}(x_i) \\ \vdots \\ 0 \end{bmatrix} \quad (2.6)
$$

을 유도할 수 있게 된다. 즉, 마지막 과정에서 계산의 편의성을 위해 $W_\lambda^\star$를 W_λ로 근사시켰으며, $S = X(W_\lambda + n\lambda I_p)^{-1}X^t$라고 두면 식 (2.6)에 의하여

$$
\begin{aligned}
\frac{\psi_\lambda(x_i) - \psi_\lambda^{(-i)}(x_i)}{y_i - \mu_\lambda^{(-i)}(x_i)} &\approx [X(W_\lambda^\star + n\lambda I_p)^{-1}X^t]_{ii} \\
&\approx [X(W_\lambda + n\lambda I_p)^{-1}X^t]_{ii} \\
&= s_{ii} (= [S]_{ii}) \qquad (2.7)
\end{aligned}
$$

를 얻게 된다. 이미 한 번 근사시킨 $CV(\lambda)$의 식 (2.4) 결과를 식 (2.7)와 결합하면, $ACV(\lambda)$ 를 얻을 수 있다.

$$
ACV(\lambda) = \frac{1}{n}l^\star + \frac{1}{n}\sum_{i=1}^{n} y_i \frac{s_{ii}(y_i - \mu_\lambda(x_i))}{1 - s_{ii}\xi''(\psi_\lambda(x_i))}
$$

$s_{ii} \approx \frac{1}{n}tr(S)$로 근사하고, $s_{ii}\xi''(\psi_\lambda(x_i)) \approx \frac{1}{n}tr(P^{\frac{1}{2}}SP^{\frac{1}{2}})$로 근사시키면 마지막으로 $GACV(\lambda)$를 얻을 수 있다.

$$GACV(\lambda) = \frac{1}{n}l^\star + \frac{tr(S)}{n}\sum_{i=1}^{n}\frac{y_i(y_i - \mu_\lambda(x_i))}{n - tr(P^{\frac{1}{2}}SP^{\frac{1}{2}})} \tag{2.8}$$

본문에서 다루게 될 자료는 이진 반응변수를 포함하기 때문에, 따라서 Bernoulli 경우를 생각하면 $\xi(\psi(x_i)) = \log(1 + e^{\psi(x_i)})$, $\xi''(\psi(x_i)) = p(x_i)(1 - p(x_i))$, $\mu_\lambda(x_i) = p_\lambda(x_i)$ 이며 P는

$$P = \begin{bmatrix} p(x_1)(1-p(x_1)) & 0 & \cdots & 0 \\ 0 & p(x_2)(1-p(x_2)) & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & p(x_n)(1-p(x_n)) \end{bmatrix}$$

이다. 따라서 최종적으로 유도한 $GACV(\lambda)$는 식 (2.9)으로 유도할 수 있음을 밝힌다.

$$GACV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[-y_i\psi_\lambda(x_i) + \xi(\psi_\lambda(x_i))] + \frac{tr(S)}{n}\sum_{i=1}^{n}\frac{y_i(y_i - p_\lambda(x_i))}{n - tr(P^{\frac{1}{2}}SP^{\frac{1}{2}})} \tag{2.9}$$

Chapter 3

Application to data

3.1. 자료 설명

본 연구는 한국거래소(이하 KRX)에서 기업 심사 후 상장폐지(Delisting) 여부를 결정짓는데 사용하는 자료를 활용하여 분석하였다. 상장폐지는 매매 대상으로서의 적격성이 없는 유가증권에 대하여 상장자격을 박탈하는 것을 의미한다. 예를들어, 유가증권시장의 상장폐지 기준(유가증권시장 상장규정 제 48조)은 KRX의 상장폐지 안내(http://listing.krx.co.kr)에서 살펴볼 수 있다. 상장폐지는 거래량, 매출액과 같은 정량적 요소와, 감사인 의견 수준(적정/부적정/한정), 상장적격성 실질심사 등과 같은 정성적 요소를 책정하여 종합적으로 평가된다. 연구에 사용한 자료는 '17년 KRX에 기 상장된 기업에 대한 정보를 내포하고 있으며, 시장 구분(유가, 코스닥 및 코넥스) 없이 전체 종목을 대상으로 분석하였다. 분석에 활용한 기 상장 종목 수는 2,583개로 두었고,

상장폐지 여부를 결정할 특징의 개수는 8개로 두었다. 특정 종목에 대한 상장폐지 여부는 투자자들에게 매우 중요한 정보이기 때문에 이와 관련한 내용은 일반적으로 공시된다. 따라서 본 연구에 사용한 기업 심사 자료는 공공성을 가지므로, 상장폐지 여부에 대한 논의가 본 논문에 가능함을 미리 밝힌다. 하지만 상장폐지 여부를 결정하는 요인으로서의 변수는 규정에 제시된 것 외에 KRX에서 사용하는 내용을 함의할 수 있으므로 구체적인 변수명은 블라인드 처리하였다. 다만, 모든 변수는 표준화하여 분석에서 생길 수 있는 불필요한 문제를 배제하였다.

3.2. 평가지표

이진 반응변수를 분류하는 모형 또는 조율모수는 다양한 지표를 이용하여 비교될 수 있다. 첫째로, 가장 단순하게 사용할 수 있는 성능지표는 정밀도(Precision)와 재현율(Recall)이다. 모형이 제시한 분류값이 실제 라벨값(상장폐지 여부)과 얼마나 일치하는지, 반대로 실제 라벨값이 모형에서 제시된 분류값과 얼마나 일치하는와 측정하여 모형의 적절성을 평가한다. 두번째 방법은 정밀도와 재현율의 가중치를 설정하여 얻을 수 있는 F_β-measure이다.

Table 3.1: 이진 분류문제의 분류 경우 표

True \Predicted	Positive	Negative
Positive	True Positive(TP)	False Negative(FN)
Negative	False Positive(FP)	True Negative(TN)

Table 3.1에서 알수 있는 정밀도, 재현율에 대한 정의는 식 (3.1), 식 (3.2)이고 , F_β measure에 대한 정의는 식 (3.3)이다. 본 자료의 Positive는 '상장폐지'를 의미한다.

$$\text{정밀도(Precision)} = \frac{\text{TP}}{\text{FP+TP}} \tag{3.1}$$

$$\text{재현율(Recall)} = \frac{\text{TP}}{\text{FN+TP}} \tag{3.2}$$

$$F_\beta - \text{measure} = \frac{(1+\beta)^2 \text{Precision} \cdot \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}} \tag{3.3}$$

식 (3.1)과 달리 정밀도(Precision)를 $\frac{\text{TN}}{\text{TN+FN}}$ 처럼 정의할 수 있을 것이다. 하지만 자료가 불균형 구조를 나타낼 경우 일반적으로 자료 개수가 더 적은 부분으로 정밀도와 재현율을 정의하고, 본 자료는 상장폐지의 경우가 매우 적은 불균형 구조이므로 식 (3.1)와 (3.2)를 이용하였다.

실제로 상장폐지에 해당하지만 실질검사에서 상장폐지가 아닌것으로 예측했다고 가정하자. 이런 유가증권이 상장된 채 시장에서 거래된다면 일반 투자자들은 건실하지 못한 기업에 투자하게 될 것이고, 그 결과 막대한 경제적 손실로 이어질 가능성이 있다. 즉, 실제 상장폐지 수준에 해당하는 기업을 상장폐지 레이블로 예측하는 것은 매우 중요하므로 False Negative가 작도록 모형을 유도해야할 것이다. 따라서 재현율의 값이 정밀도 보다 우선시 되도록 β 조정하여 $F_{1.5}$-measure를 사용하였다.

3.3. 실험 및 결과

3.3.1. 실험 방법

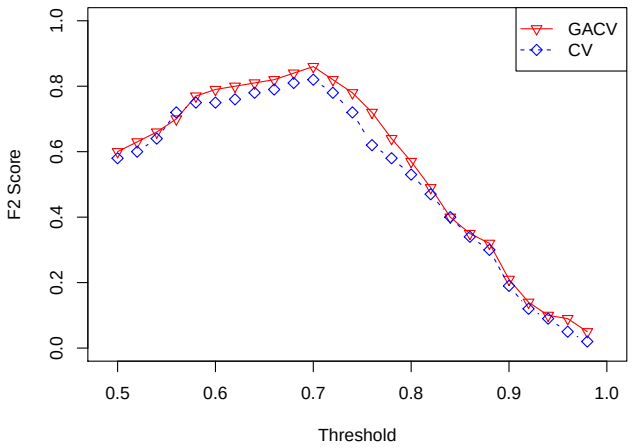
편중되지 않는 분석을 위해 전체 자료의 배열을 임의로 나열하였다. 또한 임의로 전체 자료의 55%를 훈련 자료로 두고 나머지 45%를 검증 자료로 두었다. 즉, 훈련 자료에서 *Leaving-out-one* 방법으로 교차검증방법과 일반화근사교차검증방법이 최소가 되는 λ를 얻은 뒤, 이 값을 검증 자료에 이용하여 예측하였다. 교차검증방법을 이용하면 하나의 추정치만 얻게 되는 점을 감안하여 반복적인 실험을 위해 Bootstrap을 100회 반복하였다. 따라서 전체 자료를 55:45 비율로 분할하는 작업을 100회 반복하여 각 과정에서 λ값을 여러번 구하고, 마찬가지로 검증 자료에서 정밀도, 재현율 그리고 $F_{1.5}$ 값을 구하였다.

3.3.2. 실험 결과

실험을 100회 반복하여 교차검증법과 일반화근사교차검증법으로 얻은 모형에서의 재현율, 정밀도, $F_{1.5}$ 점수를 비교해본다. 각 실험마다 얻은 평가지표들의 산술평균을 통해 검증방법을 비교하였다. 최종적으로는, 분류 라벨을 구분짓는 Threshold에 따라 재현율, 정밀도 그리고 $F_{1.5}$ 점수가 바뀌므로 Threshold를 변화시키며 모형을 비교하였다. Threshold는 0.5에서 0.98까지 0.02 간격으로 25개로 구분지어 계산하였으며 결과는 Figure 3.1으로 살펴볼 수 있다.

Threshold가 $0.5 \sim 0.6$인 구간과 0.8 이후 구간에서 GACV방법과 CV방법으로 얻은 $F_{1.5}$ 점수는 큰 차이를 보이지 않는다. 하지만 0.6에서 0.8 사이 구간에서 GACV방법으로 얻은 $F_{1.5}$ 점수는 CV방법으로 얻은 결과 값보다 근

Figure 3.1: CV, GACV $F_{1.5}$ 점수 비교



소하게 앞서는 것을 확인할 수 있다. 특히, 두 가지 교차검증방법에서 가장 큰 값의 $F_{1.5}$값을 갖는 Threshold는 모두 0.68이었다. 그 중, GACV의 $F_{1.5}$ 점수(1.363)는 CV로 얻은 $F_{1.5}$ 점수(1.321) 보다 큰 값을 나타내었다. 따라서 이진 분류를 하는 현재 모형에서 최적의 예측을 제시할 수 있는 Threshold를 0.68라고 가정하였다. 이 경우 Bootstrap을 통해 반복 교차검증을 실시한 결과 얻은 정밀도, 재현율 그리고 $F_{1.5}$ 점수는 표 3.2에 나타내었다.

Table 3.2: Threshold 0.68에서 CV, GACV의 정밀도, 재현율, $F_{1.5}$ 점수 비교

구분	Precision	Recall	$F_{1.5}$ measure
CV	0.524	0.798	1.321
GACV	0.543	0.82	1.363

Chapter 4

결론 및 제언

실험의 결과 GACV를 이용한 경우, 제시한 평가 지표 $F_{1.5}$ 점수가 CV보다 근소하게 우수한 것을 알 수 있었다. 최적의 Threshold를 탐색하기 위해 0.02 간격으로 실험하였고, CV, GACV 모두 0.68의 값에서 이진 분류 예측 모형의 $F_{1.5}$ 점수가 가장 컸음을 확인할 수 있었다. 마지막으로 최적의 Threshold 근방 구간에서 GACV의 $F_{1.5}$값이 CV의 $F_{1.5}$값보다 큰 결과를 보였다.

하지만 β값을 1.5라는 임의의 값으로 두어 정밀도와 재현율 사이의 가중치를 달리 주었는데, 이 값은 주관적인 값으로 개선의 여지가 있을 수 있다. 또한, 한 가지 평가 방법의 결과만 보고 GACV의 방법이 CV의 방법보다 우수하다 단언할 수 없다. 즉, 본 연구는 두 가지 교차검증법에 대한 우수성을 판단하기 위하여 F_β measure라는 평가 지표를 이용했지만, 정확도, 오분류율 또는 AUC 등 다양한 평가 지표들에 대한 탐색이 본 연구 자료에 시도되지 않았다. 따라서 다른 평가 지표를 이용하면 다른 결과가 나올 수 있을 가능성은 배제할 수 없다.

본 연구를 실행하며 아쉬웠던 부분은 자료 구조에 있었다. 동 자료는 구조

적으로 불균형성을 가지고 있었다. 구체적으로, 각 관측값들은 유가, 코스닥 또는 코넥스 시장에 상장되어 있는 종목이었다. 다시 말하면, 특정 유가 증권이 상장되는지 또는 상장되지 못하는지를 분류하는 문제가 아니라 기본적으로 상장된 종목 중에서 상장폐지가 되는지에 대한 분석이었다. 따라서 y 레이블에 대한 편중된 정보가 내재되었을 것이다. 불균형성을 제거하기 위해 Borderline synthetic minority over-sampling technique(SMOTE)[3] 등의 방법으로 인공 자료를 생성하였으나, 미진한 부분이 있어 본문에 적용하지 못하였다.

References

[1] D Cox and Y Chang. Iterated state space algorithms and cross validation for generalized smoothing splines. Technical report, Technical Report 49, Department of Statistics, University of Illinois, Champion, 1990.

[2] Chong Gu. Cross-validating non-gaussian data. *Journal of Computational and Graphical Statistics*, 1(2):169–179, 1992.

[3] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International Conference on Intelligent Computing*, pages 878–887. Springer, 2005.

[4] Saskia Le Cessie and Johannes C Van Houwelingen. Ridge estimators in logistic regression. *Applied statistics*, pages 191–201, 1992.

[5] Finbarr O'sullivan, Brian S Yandell, and William J Raynor Jr. Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, 81(393):96–103, 1986.

[6] Dong Xiang and Grace Wahba. A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica*, pages 675–692, 1996.

Abstract

Charm Lee Hwang
The Department of Statistics
The Graduate School
Seoul National University

In general, cross validation(CV) was applied to various statistical methodologies and used as a tool to adjust the appropriateness of the model. In addition, the generalized cross validation(GCV) was used as a tool to reduce computation efficiently in Gaussian penalized linear model with squared loss error. That is, if there is a penalty in the linear model, it is possible to appropriately select a model with high predictive power through an approximation process and a cross validation method. By extending this logic, we will examine a generalized approximate cross validation method(GACV) that generalizes the cross validation method in a penalized logistic model with binary response variables. GACV error plays a role in estimating the smoothing parameter of the regression model of the log likelihood with non-Gaussian data. The GACV can be obtained by approximating the objective function first by Taylor expansion and once again using the arithmetic mean of the smoother matrix diagonal elements as in GCV. In this paper, we show that the performance of smoothing parameters obtained by GACV using data analysis is not significantly different from the existing cross validation method.

Keywords :*Smoothing parameter, Penalized logistic regression, Cross validation, Generalized approximation cross validation.*
Student Number : 2016-20280